BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 23 2010, pages 2986–2987
doi:10.1093/bioinformatics/btq582

Structural bioinformatics

Advance Access publication October 17, 2010

FILTREST3D: discrimination of structural models using restraints from experimental data

Michal J. Gajda[1,2,†], Irina Tuszynska[1,3,†], Marta Kaczor[1], Anastasia Yu. Bakulina[4] and Janusz M. Bujnicki[1,5,*]

[1]International Institute of Molecular and Cell Biology, ul. Ks. Trojdena 4, Warsaw, Poland, [2]European Molecular Biology Laboratories, Hamburg Outstation, Notkestraße 85a, Hamburg, Germany, [3]PhD School, Institute of Biochemistry and Biophysics PAS, ul. Pawinskiego 5, Warsaw, Poland, [4]The State Research Center of Virology and Biotechnology VECTOR, Novosibirsk, Russia and [5]Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, Poznan, Poland

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Automatic methods for macromolecular structure prediction (fold recognition, *de novo* folding and docking programs) produce large sets of alternative models. These large model sets often include many native-like structures, which are often scored as false positives. Such native-like models can be more easily identified based on data from experimental analyses used as structural restraints (e.g. identification of nearby residues by cross-linking, chemical modification, site-directed mutagenesis, deuterium exchange coupled with mass spectrometry, etc.). We present a simple server for scoring and ranking of models according to their agreement with user-defined restraints.

Availability: FILTREST3D is freely available for users as a web server and standalone software at: http://filtrest3d.genesilico.pl/

Contact: iamb@genesilico.pl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 27, 2010; revised on October 7, 2010; accepted on October 12, 2010

1 INTRODUCTION

Contemporary methods for modeling of macromolecular structures are generally incapable of producing a single, well-defined, confident model and instead generate numerous alternative conformations (called 'decoys' in folding or 'poses' in docking). Benchmarking experiments CASP and CAPRI have demonstrated that among sets of alternatives there are often models that somewhat resemble the native structure; however, current algorithms still have difficulty in identifying the correct solution from the list of false positives without additional data.

It is known that the inclusion of sparse experimental data as spatial restraints can greatly improve the discrimination of near-native protein structure models from alternative conformations. Structural features of proteins are commonly studied using low-resolution methods. For example, functionally important residues that cluster together in space (such as the active sites) can be identified by mutagenesis. Exposed protein and nucleic acid surfaces or ligand-binding sites can be discovered by chemical modification. The crude topology of a protein structure can be predicted by intra- or inter-molecular cross-linking and identification of cross-linked peptide fragments by mass spectroscopy. The shape of the molecule can be studied by electron microscopy or small angle scattering of X-rays or neutrons, whereas sparse NMR data can be used to characterize local conformation of the polypeptide and identify long-range contacts. These experiments produce data that are more ambiguous, fuzzy and of much lower resolution than X-ray crystallography or fully assigned NMR. Thus, they cannot be used alone to solve the structure of a protein or a macromolecular complex. However, a combination of experimental results with bioinformatics methods significantly improves the probability of finding a correct global fold and architecture of functionally important regions (Potluri *et al.*, 2004; Ye *et al.*, 2004).

2 IMPLEMENTATION

FILTREST3D is a standalone open source Python program and a freely available web server (http://filtrest3d.genesilico.pl/) for scoring and ranking of models according to consistency with user-defined restraints (derived from experiments or computational predictions). The FILTEST3D scoring function is a simple sum of violations, expressed in real values (e.g. distance below the given threshold and solvation above a given threshold) multiplied by the weights. Weights can be assigned by the users depending on the confidence and/or accuracy of different restraints. As heterogenous experimental datasets often contain various errors, FILTREST3D allows for ranking of models based on mutually inconsistent restraints. Different restraints can also be connected with the logical operators 'and'/'or'. Discrimination with sparse or weak restraints may result in assigning good scores to many different models, which may suggest that more data should be used to allow for sharp discrimination. Alternatively, if all decoys violate the restraints, the user may consider it as a suggestion to perform additional modeling. Obviously, the interpretation of the results is the responsibility of the user.

FILTREST3D distinguishes alternative models of individual protein and nucleic acid molecules as well as protein–protein, protein–nucleic acid and protein–ligand complexes. The available restraint types include distance, solvation, local and global

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

secondary structure composition and existence of knots in protein structure. Distance restraints limit permitted range of distances between the residues (e.g. '5 Å between any atom of residue X and residue Y' or '5–10 Å between the C-α residue X in chain A and any residue of the fragment Y-Z in chain B') or sets of residues (e.g. between β-sheet A171-C182 of chain A, and an α-helix A57-D62 of chain B). Amino acid burial/exposition to the solvent may be given as a range of relative solvent accessibility (ASA) expressed in percentages. Local secondary structure can be specified for a given range of residues (e.g. 'there is an α-helix spanning residues 151–162'). Global secondary structure composition (e.g. from circular dichroism experiments) may be given as a percentage of helical and extended residues. Topological knots are rare in native structures but may occur in misfolded models; therefore, we also implemented a procedure for knot detection (Taylor, 2000). In case of mutually exclusive restraints, FILTREST3D allows different sets of mutually exclusive restraints to be applied with an 'OR' operator that returns the smallest penalty from all alternatives.

FILTREST3D generates an output file with the ranking and information about all restraints as well as a script to visualize the restraints on the model with PyMol (http://www.pymol.org). The web server automatically compares and clusters the best models, so that the user may easily discover any patterns that may emerge among multiple models that satisfy the restraints.

We tested FILTREST3D on various sets of alternative structures generated by modeling and/or docking, and tested by restraints from a variety of experimental methods. Published examples of our analyses include reinterpretation of the crystal structure of MutL C-terminal domain and identification of a correct dimer structure (Kosinski *et al*., 2005) as well as *de novo* generation of a docking complex between tRNA and two domains of methyltransferase PAB1283 (Gabant *et al*., 2006). We have also benchmarked FILTREST3D on a set of 10 CAPRI targets (Supplementary Data); our method was able to identify 657 (97%) of 678 models assessed as of high or medium quality according to CAPRI criteria.

Here, we present a practical example to discriminate between decoys obtained by low-resolution docking with GRAMM (Vakser, 1995) of a TruA enzyme structure (PDB code 1VS3, apo form) to its tRNA substrate (PDB code 2V0G, a complex with an unrelated protein). As a reference, we used the native structure of this complex (PBD code 2NR0). We produced 30 000 decoys with the grid step = 3.5 Å, and repulsion parameter 20. None of these decoys could be identified as native-like by the published potential for scoring protein–RNA interactions (Zheng *et al*., 2007), and the top-scored three decoys exhibited root mean square deviation to the native structure of >40 Å (data not shown). For discrimination of native-like complexes with FILTREST3D, we used five distance restraints. TruA residues R50 and N52 are known to be involved in the catalysis of isomerization of U39 in tRNA (two specific amino acid–nucleotide restraints) and R23, H119 and R162 are involved in RNA binding (three general restraints for interactions with any nucleotide of the whole tRNA molecule). All restraints were used with the same weights (default values). The filtering took 8 h with the standalone version of the program running on a Linux workstation with the 3.06 GHz Intel Xeon processor. Only two decoys satisfied all restraints. They were similar to each other and exhibited a native-like orientation of the tRNA with respect to the protein, despite relatively high RMSD to the native structure (22.36 and 26.56 Å; see the FILTREST3D web site for details and images). Based on this example and on our previously published analyses, we conclude that FILTREST3D allows for identification of biologically relevant models among decoys generated by low-resolution docking, even in situations where the decoys are too far from the native structure to be discriminated by other means, such as the use of potentials for protein–RNA interactions.

In principle, the same analysis can be carried out via the web server version of the program. However, the web server has a limited capacity for handling large files. Hence, the input should be split into several independent files of no more than 1 GB. Besides, the successful submission of large files is dependent on the network connection, therefore, we strongly recommend using the standalone version of FILTREST3D for scoring large sets of decoys (>1000 structures or file size >100 MB).

There exist other specialized tools for restraint-driven modeling or docking, such as MODELLER (Sali and Blundell, 1993), Haddock (Dominguez *et al*., 2003), RAPPER (Furnham *et al*., 2006) or IMP (Alber *et al*., 2007). FILTREST is not a modeling tool, but a more universal model scoring tool: it can analyze all types of models, including decoys from *de novo* folding, nucleic acids, etc. Further, to our best knowledge, FILTREST3D is the only online service that allows scoring of models based on combination of distance restraints with other factors such as local or global structure or molecule shape, and that implements logical operators to enable sets of alternative restraints.

ACKNOWLEDGEMENTS

We thank Alan Friedman and Chris Bailey-Kellogg for inspiration and discussions at the early stage of this project and Lukasz Munio for setting up the website.

Funding: Polish Ministry of Science and Higher Education (grants HISZPANIA/152/2006 and POIG.02.03.00-00-003/09); the National Institutes of Health (R01-GM081680); European Commission (LSHG-CT-2005-518238, 229676 and RIDS-011934); Human Frontier Science Program (RGP 55/2006).

Conflict of Interest: none declared.

REFERENCES

Alber,F. *et al.* (2007) Determining the architectures of macromolecular assemblies. *Nature*, **450**, 683–694.

Dominguez,C. *et al.* (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, **125**, 1731–1737.

Furnham,N. *et al.* (2006) Knowledge-based real-space explorations for low-resolution structure determination. *Structure*, **14**, 1313–1320.

Gabant,G. *et al.* (2006) THUMP from archaeal tRNA:m22G10 methyltransferase, a genuine autonomously folding domain. *Nucleic Acids Res.*, **34**, 2483–2494.

Kosinski,J. *et al.* (2005) Analysis of the quaternary structure of the MutL C-terminal domain. *J. Mol. Biol.*, **351**, 895–909.

Potluri,S. *et al.* (2004) Geometric analysis of cross-linkability for protein fold discrimination. *Pac. Symp. Biocomput.*, **9**, 447–458.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Taylor,W.R. (2000) A deeply knotted protein structure and how it might fold. *Nature*, **406**, 916–919.

Vakser,I.A. (1995) Protein docking for low-resolution structures. *Protein Eng.*, **8**, 371–377.

Ye,X. *et al.* (2004) Probabilistic cross-link analysis and experiment planning for high-throughput elucidation of protein structure. *Protein Sci.*, **13**, 3298–3313.

Zheng,S. *et al.* (2007) A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. *FEBS J.*, **274**, 6.